Liver disease Diagnosis

Tobiloba Oyediran

7/30/2021

OVERVIEW

The Indian Liver Patient Dataset contains data about liver patients in India. The data set was collected from test samples in North East of Andhra Pradesh, India. It contains records of the observed levels of certain liver-related chemical in the samples collected from patients.

Data source: https://www.kaggle.com/jeevannagaraj/indian-liver-patient-dataset (https://www.kaggle.com/jeevannagaraj/indian-liver-patient-dataset)

The goal of this project is to develop an algorithm that doctors can use to diagnose liver disease from test samples collected from patients. To achieve this, the dataset was split into training-set (80%) and validation (20%) and the training-set was further split into train-set and test-set for developing the algorithm and reviewing the model performance. The measure of performance used is the overall accuracy of the predictions.

The following steps and standard algorithms were applied to predict the outcome of the diagnosis:

1. Randomly guessing the outcome of diagnosis
2. Guessing the outcome while incorporating the observed prevalence in the dataset
3. K-nearest neighbour
4. Logistic regression
5. Linear discriminant analysis
6. Quadratic discriminant analysis
7. Random forest algorithm
8. Model ensemble

The algorithms that performed better were ensembled using majority voting to select the final prediction. There was significant improvement in the model performance (prediction accuracy increased from 0.5307692 to 0.7238095). The final ensemble algorithm was applied to the validation data. The model performance was better still (accuracy of 0.7457627).

ANALYSIS

Downloading the Indian Liver Patient Dataset

Exploration

```
## 'data.frame':    583 obs. of  11 variables:
##  $ Age                       : int  65 62 62 58 72 46 26 29 17 55 ...
##  $ Gender                    : chr  "Female" "Male" "Male" "Male" ...
##  $ Total_Bilirubin           : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.9 0.7 ...
##  $ Direct_Bilirubin          : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 0.2 ...
##  $ Alkaline_Phosphotase      : int  187 699 490 182 195 208 154 202 202 290 ...
##  $ Alamine_Aminotransferase  : int  16 64 60 14 27 19 16 14 22 53 ...
##  $ Aspartate_Aminotransferase: int  18 100 68 20 59 14 12 11 19 58 ...
##  $ Total_Protiens            : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 7.4 6.8 ...
##  $ Albumin                   : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 4.1 3.4 ...
##  $ Albumin_and_Globulin_Ratio: num  0.9 0.74 0.89 1 0.4 1.3 1 1.1 1.2 1 ...
##  $ Dataset                   : int  1 1 1 1 1 1 1 1 2 1 ...
```

Dataset contains **583 observations** of **11 variables** The long variable names were modified for convenience. The variable “Dataset” represents the diagnosis, and was modified by wrangling the data points to reflect the diagnosis.

Wrangling

```
##   age gender tot_bil direct_bil alkal_phos alam_amino aspa_amino tot_proteins
## 1  65 Female     0.7        0.1        187         16         18          6.8
## 2  62   Male    10.9        5.5        699         64        100          7.5
## 3  62   Male     7.3        4.1        490         60         68          7.0
## 4  58   Male     1.0        0.4        182         14         20          6.8
## 5  72   Male     3.9        2.0        195         27         59          7.3
## 6  46   Male     1.8        0.7        208         19         14          7.6
##   albu albu_glo_ratio diagnosis
## 1  3.3           0.90       pos
## 2  3.2           0.74       pos
## 3  3.3           0.89       pos
## 4  3.4           1.00       pos
## 5  2.4           0.40       pos
## 6  4.4           1.30       pos
```

Dataset is now in preferred format and tidy.

Any missing data?

```
## [1] TRUE
```

Some values are missing from the dataset

Which columns have missing values and how many values are missing?

```
## [1] 10 10 10 10
```

There are 4 values missing from the same column (column 10). The variable is albu_glo_ratio

Remove the observations that have missing values from the dataset.

```
## 'data.frame':    579 obs. of  11 variables:
##  $ age           : int  65 62 62 58 72 46 26 29 17 55 ...
##  $ gender        : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 1 1 2 2 ...
##  $ tot_bil       : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.9 0.7 ...
##  $ direct_bil    : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 0.2 ...
##  $ alkal_phos    : int  187 699 490 182 195 208 154 202 202 290 ...
##  $ alam_amino    : int  16 64 60 14 27 19 16 14 22 53 ...
##  $ aspa_amino    : int  18 100 68 20 59 14 12 11 19 58 ...
##  $ tot_proteins  : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 7.4 6.8 ...
##  $ albu          : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 4.1 3.4 ...
##  $ albu_glo_ratio: num  0.9 0.74 0.89 1 0.4 1.3 1 1.1 1.2 1 ...
##  $ diagnosis     : Factor w/ 2 levels "neg","pos": 2 2 2 2 2 2 2 2 1 2 ...
```

4 observations with missing albu_glo_ratio values were removed from the dataset. The cleaned dataset now has **579 observations**

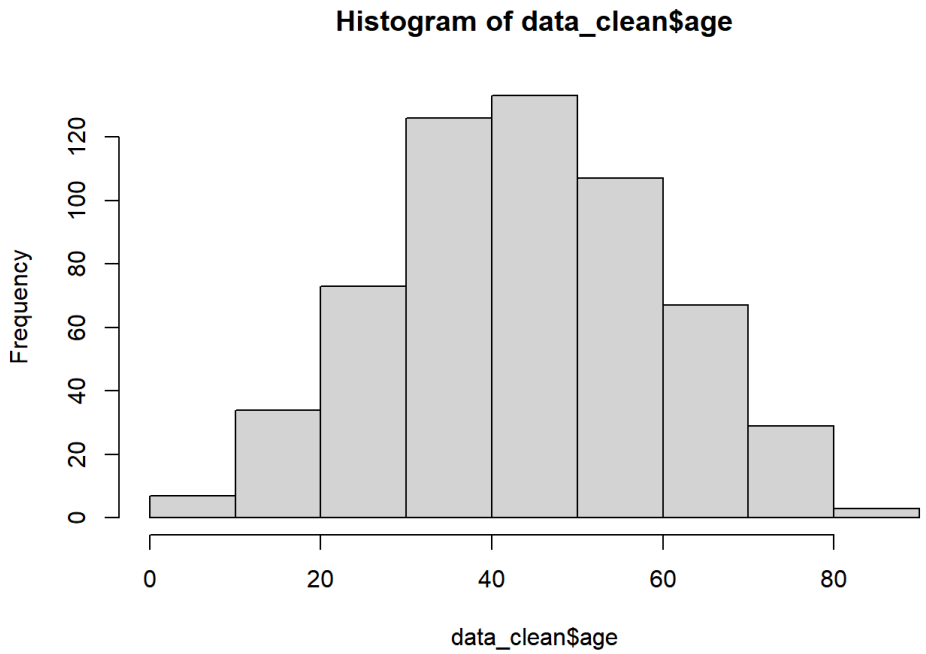
```
##   age gender tot_bil direct_bil alkal_phos alam_amino aspa_amino tot_proteins
## 1  65 Female     0.7        0.1        187         16         18          6.8
## 2  62   Male    10.9        5.5        699         64        100          7.5
## 3  62   Male     7.3        4.1        490         60         68          7.0
## 4  58   Male     1.0        0.4        182         14         20          6.8
## 5  72   Male     3.9        2.0        195         27         59          7.3
## 6  46   Male     1.8        0.7        208         19         14          7.6
##   albu albu_glo_ratio diagnosis
## 1  3.3           0.90       pos
## 2  3.2           0.74       pos
## 3  3.3           0.89       pos
## 4  3.4           1.00       pos
## 5  2.4           0.40       pos
## 6  4.4           1.30       pos
```

Dataset is now in preferred format and tidy with no missing data, ready for analysis.

Further exploration

Exploring the data further to get a grasp of the relationships (if any) that may exist among the variables in the dataset.

```
## [1]  4 90
```



Variable “age” ranges from 4 to 90

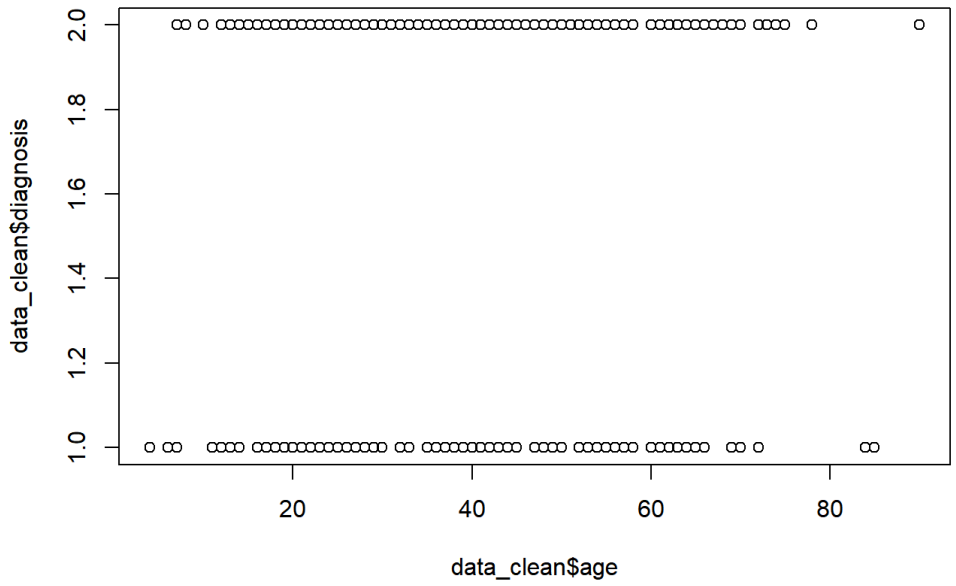
and is normally distributed throughout the dataset

How prevalent is liver disease in the dataset?

```
## [1] 0.7150259
```

About 71% of the patients in the sample have positive diagnosis of liver disease. This shows that there is a high prevalence of liver disease in the dataset.

Any relationship between age and diagnosis?

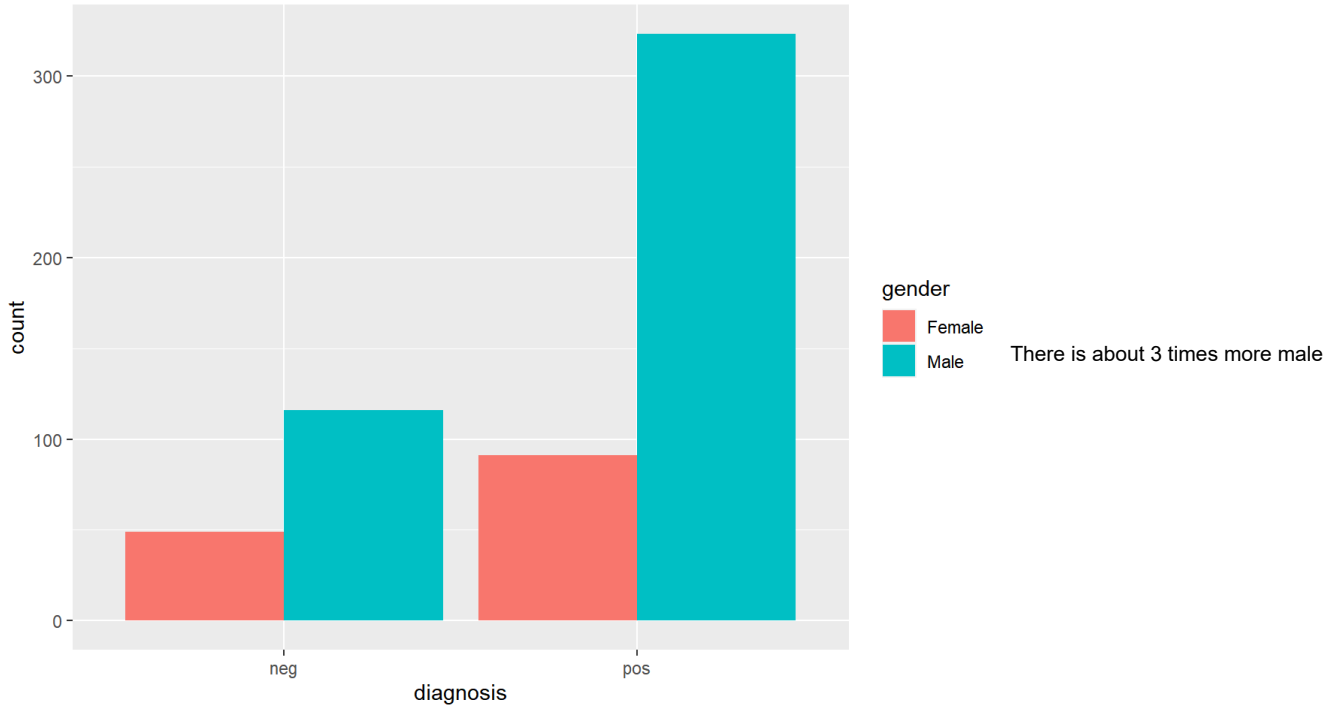


There is no observable relationship

between patient age and liver disease diagnosis

What of gender and diagnosis?

```
##
## Female   Male
##    140    439
```

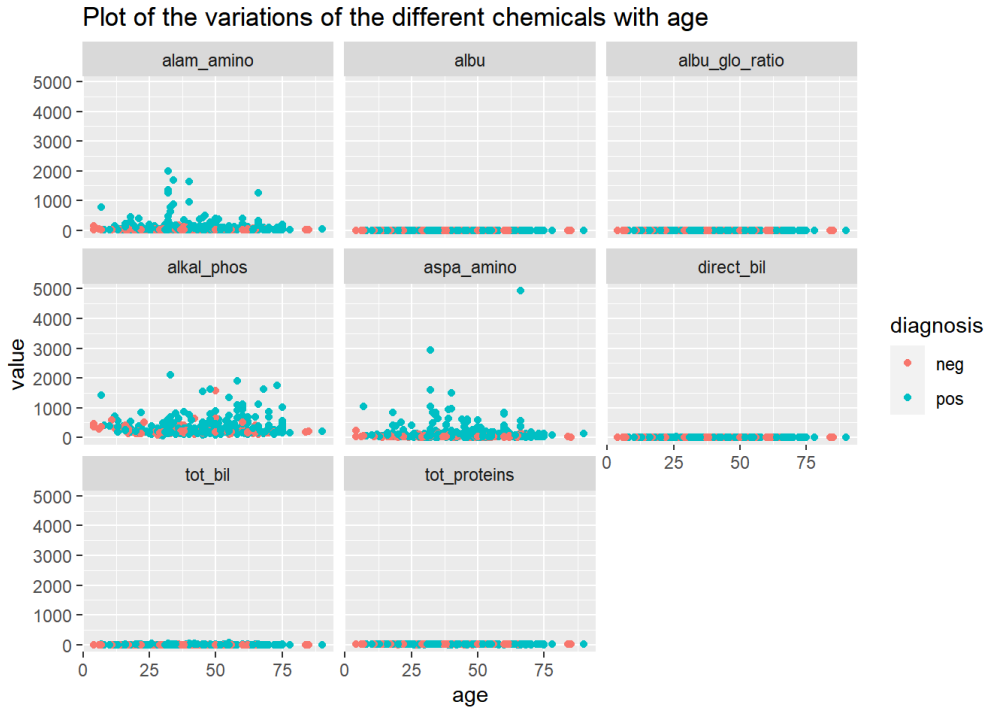


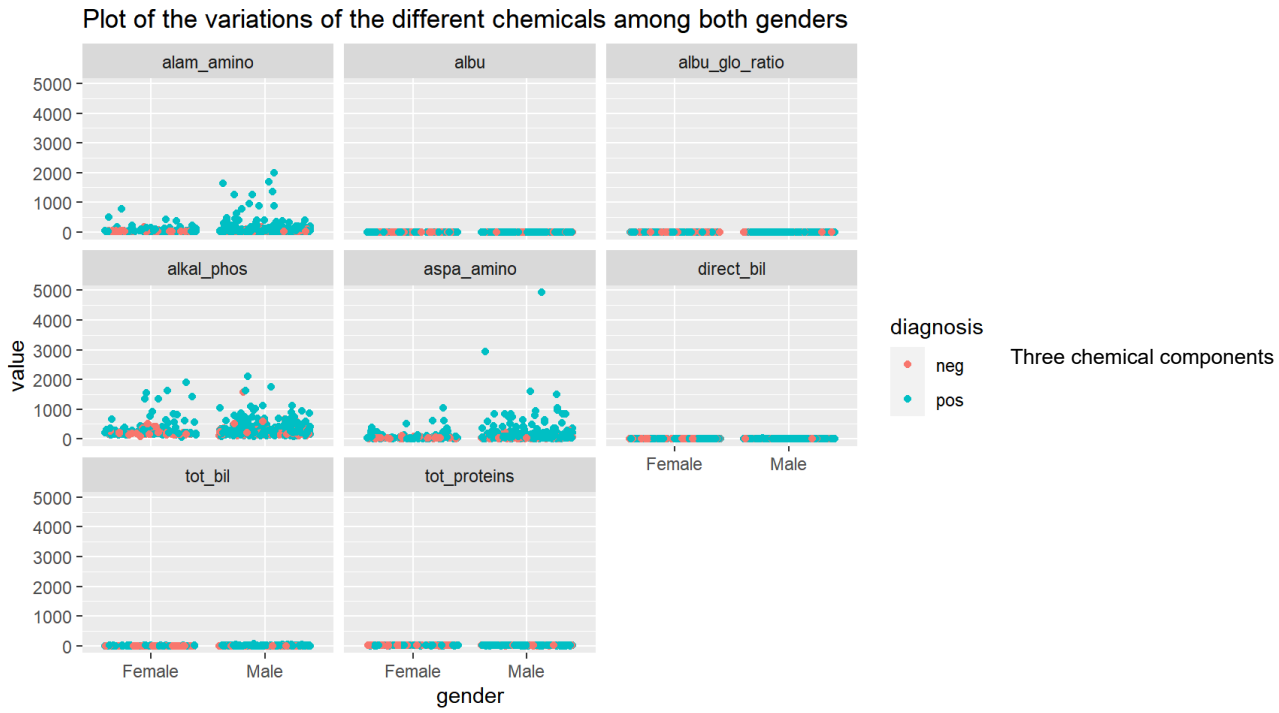
patients than female in the sample. From the plot, there seems to be a slightly higher occurrence of positive diagnosis in male patients than in female patients

Proportion of males with positive diagnosis (74%) is higher than for females (65%). Also the proportion for males is higher than the overall proportion which is 71% as seen earlier, while the proportion for female positive diagnosis is lower.

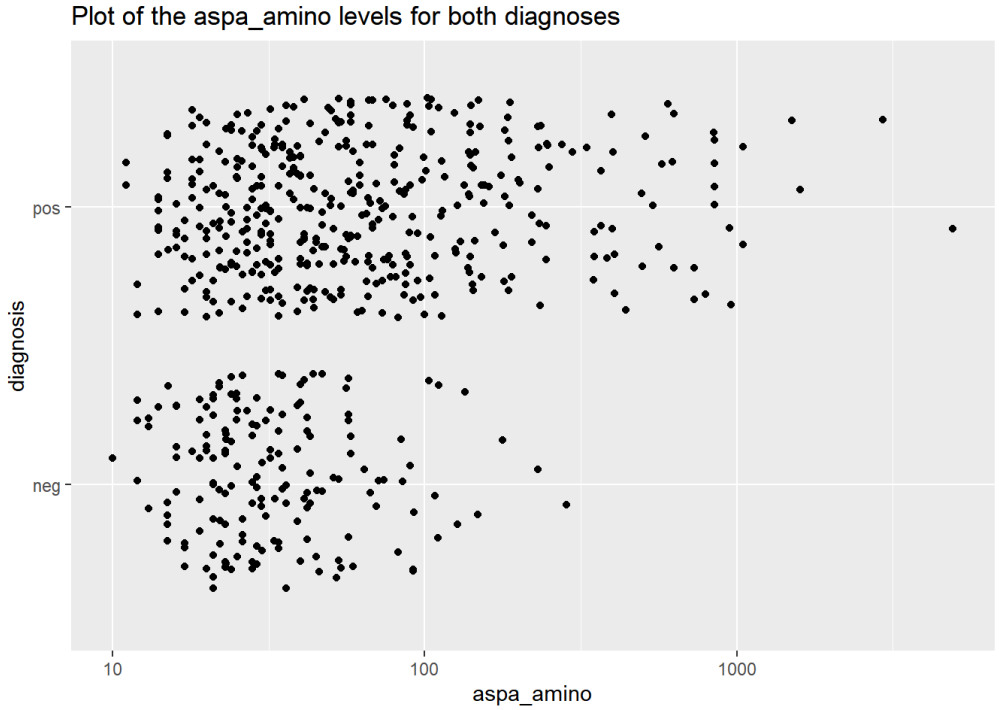
Variations is liver sample components

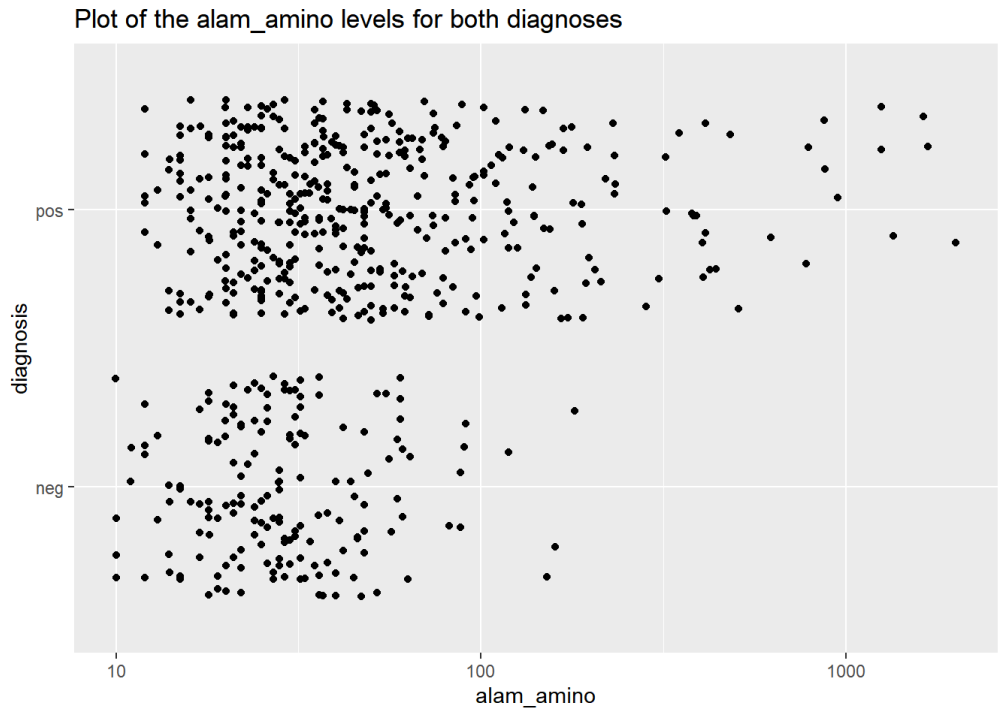
Next , let us investigate the variations in liver chemicals with age, and gender, and if they affect likelihood of liver disease

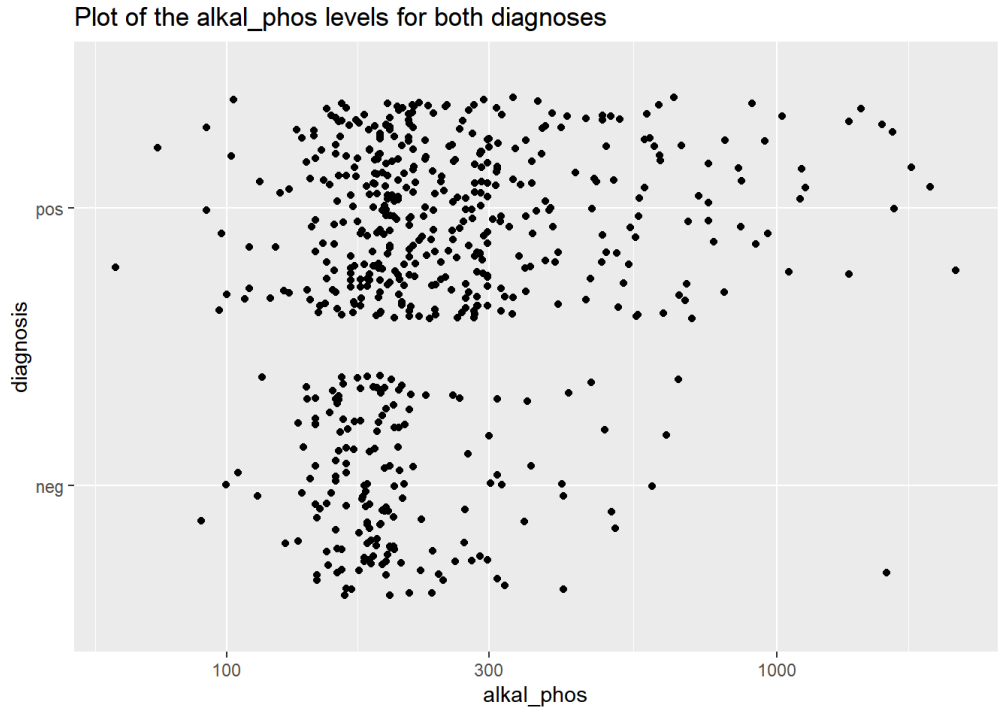




(aspa_amino, alam_amino, and alkal_phos) show some slight variation with gender and age. Let's examine them further to see if they may be possible indicators of liver disease.







From the 3 plots, the distribution of values for all 3 chemicals are similar for both positive and negative diagnoses. This means none of these chemicals can be individually used to determine the diagnosis.

ANALYSIS

Partition the data into training and validation sets

Building prediction model

Since there are no immediately apparent pointers that can be used as a starting point in building the intended classification algorithm, we can start our analysis by just randomly guessing the outcome.

Step 1: Randomly guessing the diagnosis

```
##      model Performace.rating..accuracy.
## 1 guessing                   0.5307692
```

Taking into account the higher prevalence of liver disease in the dataset, we could increase the probability of a positive diagnosis in the guess: let's use probability of 0.7

Step 2: Randomly guessing the diagnosis - with disease prevalence incorporated

```
##                     model Performace.rating..accuracy.
## 1                guessing                    0.5307692
## 2 guessing with prevalence                    0.5980769
```

With disease prevalence incorporated, we see some improvement in model performance.

Now let's apply some of the standard machine learning models to the data and see if we will get any better performance for this model.

Partition the training data into and train and test sets

Step 3: Applying K-nearest neighbours algorithm for diagnosis prediction

```
##                           model Performace.rating..accuracy.
## 1                      guessing                    0.5307692
## 2         guessing with prevalence                    0.5980769
## 3 K - nearest neighbours algorithm                    0.6666667
```

We see some improvement in performance beyond the improved guessing model

**Steps 4-7*: Now let's apply other standard machine learning algorithms to generate predictions and see how they all perform

```
##                           model Performace.rating..accuracy.
## 1                      guessing                    0.5307692
## 2         guessing with prevalence                    0.5980769
## 3 K - nearest neighbours algorithm                    0.6666667
## 4              Logistic regression                    0.7523810
## 5     Linear discriminant analysis                    0.7142857
## 6  Quadratic discriminant analysis                    0.5619048
## 7                   Random forest                    0.7428571
```

From the table of model performance results, we see that all the standard ML models, except the Quadratic discriminant analysis (QDA), do better than our improved guessing model. The best performing model is the Logistic regression model, followed by the Random forest algorithm, and then the Linear discriminant analysis model.

Step 8: Creating an Ensemble

Since the QDA algorithm preformed even worse than the improved guessing model, we can ignore the QDA algorithm. We can also ignore the KNN algorithm as it does not perform so much better than the improved guessing model. So we include only the 3 models that have a performance rating of 0.7 and above (i.e. Logistic regression, LDA, and Random forest) in the ensemble and use majority voting method to select the final prediction

```
##                           model Performace.rating..accuracy.
## 1                      guessing                    0.5307692
## 2         guessing with prevalence                    0.5980769
## 3 K - nearest neighbours algorithm                    0.6666667
## 4              Logistic regression                    0.7523810
## 5     Linear discriminant analysis                    0.7142857
## 6  Quadratic discriminant analysis                    0.5619048
## 7                   Random forest                    0.7428571
## 8                  Model ensemble                    0.7428571
```

The performance of the model ensemble is better than that of the LDA, at par with Random forest, but not as good as the Logistic regression model.

We can now apply this model ensemble to predict diagnosis in the validation data. This will be the final test of the performance of the model we have built so far.

Final Evaluation

Using the whole training-set to train the algorithm and validation as the test set

```
## [1] "Final model ensemble performance rating is  0.7457627"
```

Let's compare this final performance level with what could have been obtained if we simply used the standard model with the best performance on the training data (the Logistic regression model),

```
## [1] "Final Logistic regression model performance rating is  0.7457627"
```

RESULTS

From the results table, it is evident that:

1. Incorporating the idea of disease prevalence in the sample produced better performance while randomly guessing the diagnosis. This shows that even a little understanding of the dataset can help to improve predictions.

2. There is incremental improvement in performance of the model when standard machine learning models were applied to generate predictions from the training data.
3. Although the model ensemble did not perform as well as the Logistic regression model on the training data, it performed equally well when applied to the validation data.
4. Comparing the performance of the model ensemble on the test_set and validation data, the model performed better when a larger training set is used (training_set is larger than train_set). This suggests that the larger the training dataset, the lower the error of prediction using this model.

CONCLUSION

The model performance has been significantly improved by creating an ensemble of the best performing standard machine learning models.

Limitation of this work

This report does not explore the possibility of creating an algorithm that could predict diagnosis based on slight changes in the level of chemical components. There could be particular level of the chemical components beyond which diagnosis can be accurately predicted as either positive or negative. Also a mixture or combination of some of the variables in certain ways could produce direct pointers to accurate diagnosis.

Future work

To further improve the performance of the prediction model, principal component analysis could be used to explore the possibility of discovering indicators that can directly point to accurate diagnosis of liver disease.